Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 3 —

Ali Shakiba

Vali-e-Asr University of Rafsanjan

based on slides by

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign

Simon Fraser University

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview

    - Data Quality

    - Major Tasks in Data Preprocessing

- Data Cleaning

- Data Integration

- Data Reduction

- Data Transformation and Data Discretization

- Summary

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view

    - Accuracy: correct or wrong, accurate or not

    - Completeness: not recorded, unavailable, …

    - Consistency: some modified but some not, dangling, …

    - Timeliness: timely update?

    - Believability: how trustable the data are correct?

    - Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview

  - Data Quality

  - Major Tasks in Data Preprocessing

- Data Cleaning

- Data Integration

- Data Reduction

- Data Transformation and Data Discretization

- Summary

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error

  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

    - e.g., *Occupation*=" " (missing data)

  - noisy: containing noise, errors, or outliers

    - e.g., *Salary*="−10" (an error)

  - inconsistent: containing discrepancies in codes or names, e.g.,

    - *Age*="42", *Birthday*="03/07/2010"

    - Was rating "1, 2, 3", now rating "A, B, C"

    - discrepancy between duplicate records

  - Intentional (e.g., *disguised missing* data)

    - Jan. 1 as everyone's birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably

- Fill in the missing value manually: tedious + infeasible?

- Fill in it automatically with

  - a global constant : e.g., "unknown", a new class?!

  - the attribute mean

  - the attribute mean for all samples belonging to the same class: smarter

  - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- **Noise**: random error or variance in a measured variable
- **Incorrect attribute values** may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- **Other data problems** which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- **Binning**
    - first sort data and partition into (equal-frequency) bins
    - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- **Regression**
    - smooth by fitting the data into regression functions
- **Clustering**
    - detect and remove outliers
- **Combined computer and human inspection**
    - detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning as a Process

- Data discrepancy detection
    - Use metadata (e.g., domain, range, dependency, distribution)
    - Check field overloading
    - Check uniqueness rule, consecutive rule and null rule
    - Use commercial tools
        - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
        - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
    - Data migration tools: allow transformations to be specified
    - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
    - Iterative and interactive (e.g., Potter's Wheels)